# EXHIBIT B

**UNITED STATES DISTRICT COURT**
**SOUTHERN DISTRICT OF NEW YORK**

| | |
|---|---|
| AUTHORS GUILD, et al.,<br><br>Plaintiffs,<br><br>v.<br><br>OPEN AI INC., et al.,<br><br>Defendants. | **ECF CASE**<br><br>No. 1:23-cv-08292-SHS;<br>No. 1:23-cv-10211-SHS<br><br>**PLAINTIFFS' THIRD SET OF REQUESTS FOR PRODUCTION TO OPENAI** |
| JONATHAN ALTER, et al.,<br><br>Plaintiffs,<br><br>v.<br><br>OPENAI, INC., et al.,<br><br>Defendants. | |

**PLAINTIFFS' THIRD SET OF REQUESTS FOR PRODUCTION TO OPENAI**

Pursuant to Rules 26 and 34 of the Federal Rules of Civil Procedure, Plaintiffs request that Defendant OpenAI respond to the following Third Set of Requests for Production of Documents ("Requests"). Responses to these Requests shall be due within thirty (30) days of the date of service or as otherwise mutually agreed by the parties. Plaintiffs are amenable to an electronic production, subject to agreement by the parties.

In accordance with Rule 34(b), OpenAI shall provide written responses to the following Requests and shall produce the requested documents as they are kept in the ordinary and usual course of business or shall organize and label the documents to correspond with the categories in these Requests. In accordance with Rule 26(e), OpenAI shall supplement or correct its responses

or productions as necessary.

## I. DEFINITIONS

1. **"Actions"** means the above captioned litigation, *Authors Guild et al. v. Open AI Inc. et al.*, No. 1:23-cv-08292-SHS (S.D.N.Y.) (opened Sept. 19, 2023 and amended December 5, 2023), and *Alter et al. v. OpenAI et al.*, No / 1:23-cv-10211-SHS (S.D.N.Y.) (opened Nov. 21, 2023, and amended Dec. 19, 2023).

2. **"Communication(s)"** means the transmittal of information (in the form of facts, ideas, inquiries or otherwise) by any means, including, but not limited to, telephone calls, emails (whether via company server or personal webmail or similar accounts), faxes, text messages (on work or personal phones), instant messages, Skype, Line, WhatsApp, WeChat, other electronic messages, letters, notes, and voicemails.

3. **"Document(s)"** is defined to be synonymous in meaning and equal in scope to the usage of the term "documents or electronically stored information" in Rule 34(a)(1)(A). For the avoidance of doubt, **Document(s)** includes **Communication(s)**.

4. **"You", "Your"**, and **"OpenAI"** means OpenAI, Inc., OpenAI GP, LLC, OpenAI, LLC, OpenAI OPCO LLC, OpenAI Global LLC, OAI Corporation, LLC, OpenAI Holdings, LLC, and any of their directors, officers, employees, partners, members, representatives, agents (including attorneys, accountants, consultants, investment advisors or bankers), and any other person acting or purporting to act on their behalf, as well as corporate parents, subsidiaries, affiliates, predecessor entities, successor entities, divisions, departments, groups, acquired entities, related entities, or any other entity acting or purporting to act on their behalf.

5. **"Large Language Model," "LLM," "AI Model(s)," "Generative AI system(s)," "model(s),"** and **"API Product(s)"** have the same meaning as they are used in YOUR letter to the Register of Copyrights and Director of the U.S. Copyright Office dated October 30, 2023, "Re:

Notice of Inquiry and Request for Comment [Docket No. 2023-06]" and include all models listed or described in https://platform.openai.com/docs/models.

6. **"Person(s)"** means any individual or entity.

7. **"Fine Tune(d)," "Fine Tuning," "Pre-Train(ed)," "Pre-Training," "Train[ed],"** and **"Training,"** have the same meaning as the term is discussed in **Your** website materials and statements. See e.g., https://platform.openai.com/docs/guides/fine-tuning ("OpenAI's text generation models have been pre-trained on a vast amount of text. To use the models effectively, we include instructions and sometimes several examples in a prompt. Using demonstrations to show how to perform a task is often called "few-shot learning." Fine-tuning improves on few-shot learning by training on many more examples than can fit in the prompt, letting you achieve better results on a wide number of tasks. Once a model has been fine-tuned, you won't need to provide as many examples in the prompt."); https://openai.com/index/language-unsupervised/ ("These results provide a convincing example that pairing supervised learning methods with unsupervised pre-training works very well; this is an idea that many have explored in the past, and we hope our result motivates further research into applying this idea on larger and more diverse datasets."); and https://openai.com/index/how-should-ai-systems-behave/ ("The two main steps involved in building ChatGPT work as follows . . . First, we "pre-train" models by having them predict what comes next in a big dataset . . . Then, we "fine-tune" these models on a more narrow dataset that we carefully generate with human reviewers who follow guidelines that we provide them."); https://openai.com/index/gpt-4-research/ ("Interestingly, the base pre-trained model is highly calibrated . . . Note that the model's capabilities seem to come primarily from the pre-training process . . . "); https://openai.com/index/text-and-code-embeddings-by-contrastive-pre-training/ ("In this work, we show that contrastive pre-training on unsupervised data at scale leads to high

quality vector representations of text and code."); https://openai.com/index/vpt/ ("We trained a neural network to play Minecraft by Video PreTraining (VPT) on a massive unlabeled video dataset of human Minecraft play, while using only a small amount of labeled contractor data. With fine-tuning, our model can learn to craft diamond tools . . .").

8. **"ChatGPT"** means all consumer-facing versions of the chatbot application OpenAI released in November 2022 and any **LLM** underlying those consumer-facing applications.

9. **"Concern"** means be the subject of, make reference to, comment on, discuss, describe, identify, or contain text or images about the stated topic.

10. **"Including"** means including but not limited to.

11. **"Investors"** means any person or entity that financially invested in OpenAI in the operations and/or entitie(s) controlled and/or overseen by one or more of the OpenAI defendant entities.

12. **"Microsoft"** means Microsoft Corporation and any of its directors, officers, employees, partners, members, representatives, agents (INCLUDING attorneys, accountants, consultants, investment advisors or bankers), and any other person acting or purporting to act on their behalf, as well as corporate parents, subsidiaries, affiliates, predecessor entities, successor entities, divisions, departments, groups, acquired entities, related entities, or any other entity acting or purporting to act on its behalf.

13. **"Relate to"** means refer to, respond to, describe, evidence, or constitute, in whole or in part.

14. **"Publishers**" means a person or entity whose business is publishing materials, including, but not limited to, books, periodicals, magazines, and newspapers.

## II. RELEVANT TIME PERIOD

The relevant time period is January 1, 2015 through the present ("Relevant Time Period"),

unless otherwise specifically indicated, and shall include all **Documents** and any other information relating to such period, even though prepared or published outside of the Relevant Time Period. If a **Document** prepared before the Relevant Time Period is necessary for a correct or complete understanding of any **Document** covered by any of these Requests, please provide the earlier **Document** as well. If any **Document** is undated and the date of its preparation cannot be determined, please produce the **Document** if it is otherwise responsive to any Request.

## III. INSTRUCTIONS

1. The production by one person, party, or entity of a **Document** does not relieve another person, party, or entity from the obligation to produce his, her, or its own copy of that **Document**.

2. Produce **Documents** not otherwise responsive to these Requests if such **Documents Relate to** the **Documents** that are called for by these Requests, or if such **Documents** are attached to **Documents** called for by these Requests.

3. Produce each **Document** requested herein in its entirety and without deletion, excisions, redaction, or other modification regardless of whether **You** consider the entire document to be relevant or responsive.

4. If any **Document** is known to have existed but no longer exists, has been destroyed, or is otherwise unavailable, identify the **Document**, the reason for its loss, destruction, or unavailability, the name of each person known or reasonably believed by **Microsoft** to have had possession, custody, or control of the original and any copy thereof (if applicable), and a description of the disposition of each copy of the **Document**.

5. If no **Documents** responsive to a particular Request exist, state that no responsive **Documents** exist.

6. If You assert that any of the **Documents** and things requested are protected from discovery by attorney-client privilege, the attorney work product doctrine, or any other evidentiary privilege, specify for each **Document** (1) the grounds asserted as the reason for non-production; (2) the date the **Document** was prepared; (3) the identity of the attorney(s) who drafted or received the **Document(s)** (if attorney-client privilege or attorney work product protection is claimed); (4) the identity of the parties who prepared or received the **Document**; and (5) the nature of the **Document**.

7. Construe the conjunctions "and" and "or" non-restrictively or non-exclusively if doing so would bring within the scope of these Requests **Documents** that might otherwise be construed to be outside of their scope.

8. Construe the use of the singular to include the plural; the use of the masculine, feminine, or neuter gender to include the others; and the use of one form of the verb to include the others if doing so would bring within the scope of these Requests **Documents** that might otherwise be construed to be outside of their scope.

Plaintiffs, by and through their undersigned attorneys, request that Defendants provide a response to these Requests within thirty (30) days of the date of service hereof as provided by Federal Rule of Civil Procedure 33.

## REQUESTS FOR PRODUCTION

**REQUEST FOR PRODUCTION NO. 33:**

All **Documents** regarding the deletion of any training datasets, including but not limited to Books1 and/or Books2 and including but not limited to **Documents** sufficient to identify such deleted datasets, the type of materials contained or believed to have been contained in those training datasets, and the date of deletion.

**REQUEST FOR PRODUCTION NO. 34:**

Retention agreements between **You** and any law firm(s) for matters related to copyright litigation or potential copyright litigation related your development and/or of any artificial intelligence models.

**REQUEST FOR PRODUCTION NO. 35:**

**Your** retention letters with Morrison & Foerster LLP and/or Latham & Watkins LLP executed between January 1, 2016 and December 31, 2022.

**REQUEST FOR PRODUCTION NO. 36:**

All **Communications** with third parties between January 1, 2016 and December 31, 2022 related to the use or reproduction of copyrighted materials or allegedly copyrighted materials to train any artificial intelligence model **You** developed, including without limitation all large language models and diffusion models (e.g., DALL-E, Codex, GPT-2).

**REQUEST FOR PRODUCTION NO. 37:**

All **Documents Concerning** or **Relating to** the preparation of the comment **You** submitted in response to the U.S. Patent and Trademark Office's Request for Comments on Intellectual Property Protection for Artificial Intelligence Innovation issued on October 30, 2019, including without limitation all **Documents** provided to the people who drafted that comment.

**REQUEST FOR PRODUCTION NO. 38:**

All **Communications** between **You** and any **Person**, including but not limited to **Publishers**, from whom you obtained or sought to obtain a license for the use of text or audio in connection with **Large Language Models**.

**REQUEST FOR PRODUCTION NO. 39:**

All **Documents Concerning** or **Relating to** curation and the type of data used to train **Large Language Models**, including but not limited to **Documents Concerning** or **Relating to** the importance of long stretches of text, of high-quality text, or of human-written text; differences in the type or quality of output produced by models that do or do not contain different kinds of data; differences in the way in which **Large Language Models** or **ChatGPT** ingest, process, or output such data; and the decision to include or exclude data from the training data of certain **Large Language Models**.

**REQUEST FOR PRODUCTION NO. 40:**

**Documents Concerning** or **Relating to** the relative value of types of data for use in training Large Language Models.

**REQUEST FOR PRODUCTION NO. 41:**

To the extent not already requested, **Documents Concerning** or **Relating to** modifying any parameters for **Tuning** or restricting the output of **Large Language Models** or **ChatGPT** designed or intended—in whole or in part—to avoid copyright infringement and/or to avoid reproducing materials used to train the Large Language Models or ChatGPT.

**REQUEST FOR PRODUCTION NO. 42:**

**Documents Concerning** or **Relating to Your** creating human-generated text for use as training data for **Large Language Models**.

**REQUEST FOR PRODUCTION NO. 43:**

**All Documents Concerning** or **Relating to** changes in output quality of **Large Language Models** based on the inclusion or exclusion of human-generated text in the training data for **Large Language Models**.

**REQUEST FOR PRODUCTION NO. 44:**

All Documents Concerning or Relating to ChatGPT or Large Language Models commercial uses and/or applications.

**REQUEST FOR PRODUCTION NO. 45:**

All **Documents** Concerning or Relating to Your potential or actual conversion to a for-profit entity.

**REQUEST FOR PRODUCTION NO. 46:**

All **Documents Concerning** or **Relating to** whether **Your** use of copyrighted material in **Large Language Models** complies with copyright law in the United States or elsewhere.

**REQUEST FOR PRODUCTION NO. 47:**

**All Documents Concerning or Relating to** products, services or "in-kind" investments provided to **You** by **Microsoft**, including but not limited to the need for or importance of such products, services, or "in-kind" investments.

**REQUEST FOR PRODUCTION NO. 48:**

All **Documents Concerning** or **Relating to** the use or inclusion of **ChatGPT** or **Your Large Language Models** in **Microsoft's** products, including but not limited to Azure, and including but not limited to **Documents Concerning** or **Relating to** any profits or revenues from such products and the distribution thereof among **You** and **Microsoft**.

**REQUEST FOR PRODUCTION NO. 49:**

All **Documents Concerning** or **Relating to** the use or inclusion of **ChatGPT** or **Your Large Language Models** in commercial products, including but not limited to **Documents Concerning** or **Relating to** the ideation, development, and commercialization timeframe.

**REQUEST FOR PRODUCTION NO. 50:**

For each **Large Language Model You** have commercialized, sold, and/or licensed, the gross revenues, net revenues, and profits, by month, generated by the **Large Language Model**.

**REQUEST FOR PRODUCTION NO. 51:**

The balance sheet, income statement, and cash flow statement, on a monthly basis, for each of OpenAI, Inc., OpenAI GP, LLC, OpenAI, LLC, OpenAI OPCO LLC, OpenAI Global LLC, OAI Corporation, LLC, and OpenAI Holdings, LLC.

**REQUEST FOR PRODUCTION NO. 52:**

All **Documents** sufficient to show the output of any **Large Language Models** created by **You** related to Plaintiffs' works.

**REQUEST FOR PRODUCTION NO. 53:**

**Documents** sufficient to show each version of the terms and/or conditions of use for **ChatGPT**.

**REQUEST FOR PRODUCTION NO. 54:**

**All Documents Concerning** or **Relating to Your** decision to include, modify, or remove the following or similar language from **Your** terms and/or conditions of use for **ChatGPT**: "With respect to the content or other materials you upload through the Site or share with other users or recipients (collectively, "User Content"), you represent and warrant that you own all right, title and interest in and to such User Content, including, without limitation, all copyrights and rights of publicity contained therein. By uploading any User Content you hereby grant and will grant OpenAI and its affiliated companies a nonexclusive, worldwide, royalty free, fully paid up, transferable, sublicensable, perpetual, irrevocable license to copy, display, upload, perform,

distribute, store, modify and otherwise use your User Content for any OpenAI-related purpose in any form, medium or technology now known or later developed."

**REQUEST FOR PRODUCTION NO. 55:**

All **Documents Concerning** or **Relating to** any version of Library Genesis, LibGen, Internet Archive, Z-Library, Common Crawl, WebText, Project Gutenberg, Anna's Archive, Open Library, Reddit, DuXiu, or any other repository of text You used or considered using to train Your Large Language Models.

**REQUEST FOR PRODUCTION NO. 56:**

All **Documents Concerning** or **Relating to** the paper entitled "Fair Learning" by Mark Lemley and Bryan Casey.

**REQUEST FOR PRODUCTION NO. 57:**

**All Documents Concerning** or **Relating to** the impact or potential future impact of Your **Large Language Models, ChatGPT,** or artificial intelligence on the markets for labor, art, or writing.

**REQUEST FOR PRODUCTION NO. 58:**

To the extent not already requested, **Documents Concerning** or **Relating to Your** plans to allow authors, creators, or rightsholders to either receive compensation or determine how and whether their works will be used by You, including All **Documents Concerning** or **Relating to** Sam Altman's congressional testimony that "creators deserve control over how their creations are used".

**REQUEST FOR PRODUCTION NO. 59:**

All **Documents Concerning or Relating To** Your decision(s) to obtain or request a license for training material from any copyright holder.

**REQUEST FOR PRODUCTION NO. 60:**

**All Documents** describing the way in which **Your Large Language Models** use training data, including but not limited to any metaphor, comparison, or analogy between likening (or rejecting the likening of) **Large Language Model** training to (a) human learning, (b) pattern recognition, or (c) autocomplete/parroting.

**REQUEST FOR PRODUCTION NO. 61:**

**Documents** sufficient to show all audiobooks You transcribed.

**REQUEST FOR PRODUCTION NO. 62:**

To the extent not already requested, **Documents Concerning** or **Relating to** the exhaustion of data available for training a Large Language Model.

**REQUEST FOR PRODUCTION NO. 63:**

To the extent not already requested, all **Documents Concerning** or **Relating to Your** decision to stop making public information about the training of **Your Large Language Models**.

**REQUEST FOR PRODUCTION NO. 64:**

All **Documents**, **Communications**, and tangible things that **You** have in **Your** possession, custody, or control and may use to support **Your** claims and/or defenses, to oppose or support any motion in this Action, and/or as evidence or for impeachment at trial.

**REQUEST FOR PRODUCTION NO. 65:**

**All Documents Concerning** or **Relating to Your** use of web crawler permissions in connection with your efforts to comply with copyright law, including **Documents Concerning** or **Relating to Your** awareness, belief, understanding, or speculation that the use of web crawler permissions is not sufficient to exclude copyrighted materials from the training data.

**REQUEST FOR PRODUCTION NO. 66:**

**All Documents Concerning** or **Relating to** Media Manager, including **Documents** sufficient to show all possible features of Media Manager discussed by **You**.

**REQUEST FOR PRODUCTION NO. 67:**

**Documents** sufficient to show all prompts inputted into ChatGPT relating to or concerning any of the Class Works authored by Named Plaintiffs and the output ChatGPT generated in response to each prompt.

**REQUEST FOR PRODUCTION NO. 68:**

All **Documents** produced by **You** in *Tremblay v. OpenAI, Inc.*, No. 23-cv-3223 (N.D. Cal.).

**REQUEST FOR PRODUCTION NO. 69:**

All **Documents Concerning** or **Relating to** whether **Your Large Language Models** comprise a compression of training data.

**REQUEST FOR PRODUCTION NO. 70:**

All **Documents Concerning** or **Relating to** the dataset known as The Pile and/or The Eye.

Dated: June 14, 2024

*/s/ Justin A. Nelson*
Justin A. Nelson (*pro hac vice*)
Alejandra C. Salinas (*pro hac vice*)
SUSMAN GODFREY L.L.P.
1000 Louisiana Street, Suite 5100
Houston, TX 77002
Tel.: 713-651-9366
jnelson@susmangodfrey.com
asalinas@susmangodfrey.com

Rohit D. Nath (*pro hac vice*)
SUSMAN GODFREY L.L.P.
1900 Avenue of the Stars, Suite 1400
Los Angeles, California 90067
Tel.: 310-789-3100

rnath@susmangodfrey.com

J. Craig Smyser
SUSMAN GODFREY L.L.P.
1901 Avenue of the Americas, 32nd Floor
New York, New York 10019
Tel.: 212-336-8330
csmyser@susmangodfrey.com

*/s/ Rachel Geman*
Rachel Geman
LIEFF CABRASER HEIMANN & BERNSTEIN, LLP
250 Hudson Street, 8th Floor
New York, New York 10013-1413
Tel.: 212-355-9500
rgeman@lchb.com

Reilly T. Stoler (*pro hac vice forthcoming*)
LIEFF CABRASER HEIMANN & BERNSTEIN, LLP
275 Battery Street, 29th Floor
San Francisco, CA 94111-3339
Tel.: 415-956-1000
rstoler@lchb.com
ibenserg@lchb.com

Wesley Dozier (*pro hac vice*)
LIEFF CABRASER HEIMANN & BERNSTEIN, LLP
222 2nd Avenue, Suite 1640
Nashville, TN 37201
Tel.: 615-313-9000
wdozier@lchb.com

*/s/ Scott J. Sholder*
Scott J. Sholder
CeCe M. Cole
COWAN DEBAETS ABRAHAMS & SHEPPARD LLP
41 Madison Avenue, 38th Floor
New York, New York 10010
Tel.: 212-974-7474
ssholder@cdas.com
ccole@cdas@com
***Attorney for Plaintiffs and the Proposed Class***

**CERTIFICATE OF SERVICE**

I hereby certify that on June 14, 2024, a copy of the foregoing was served via electronic mail to all counsel of record in this matter.

*/s/ Ellen Sullivan-Vasquez*
Ellen Sullivan-Vasquez